

Distributed optimization over time-varying directed graphs

Angelia Nedić and Alex Olshevsky*

March 13, 2013

Abstract

We consider distributed optimization by a collection of nodes, each having access to its own convex function, whose collective goal is to minimize the sum of the functions. The communications between nodes are described by a time-varying sequence of *directed* graphs, which is uniformly strongly connected. For such communications, assuming that every node knows its out-degree, we develop a broadcast-based algorithm, termed the *subgradient-push*, which steers every node to an optimal value under a standard assumption of subgradient boundedness. The subgradient-push requires no knowledge of either the number of agents or the graph sequence to implement. Our analysis shows that the subgradient-push algorithm converges at a rate of $O(\ln t/\sqrt{t})$, where the constant depends on the initial values at the nodes, the subgradient norms, and, more interestingly, on both the consensus speed and the imbalances of influence among the nodes.

1 Introduction

We consider the problem of distributed convex optimization by a network of nodes when knowledge of the objective function is scattered throughout the network and unavailable at any single location. There has been much recent interest in multi-agent optimization problems of this type that arise whenever a large collections of nodes - which may be processors, nodes of a sensor network, vehicles, or UAVs - desire to collectively optimize a global objective by means of local actions taken by each node without any centralized coordination.

Specifically, we will study the problem of optimizing a sum of n convex functions by a network of n nodes when each function is known to only a single node. This problem frequently arises when control and signal processing protocols need to be implemented in sensor networks. For example, the problems including robust statistical inference [22], formation control [20], non-autonomous power control [25], distributed message routing [19],

*The authors are with the Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, 104 S. Mathews Avenue, Urbana IL, 61801, Emails: {angelia,aolshev2}@illinois.edu. A. Nedić gratefully acknowledges the support by the NSF under grants CMMI 07-42538 and CCF 11-11342.

and spectrum access coordination [10], can be reduced to variations of this problem. We will be focusing on the case when communication between nodes is *directed* and *time-varying*.

Distributed optimization of a sum of convex functions has received a surge of interest in recent years [17, 22, 8, 13, 11, 12, 27, 15, 3, 24, 5]. There is now a considerable theory justifying the use of distributed subgradient methods in this setting, and their performance limitations and convergence times are well-understood. Moreover, distributed subgradient methods have been used to propose new solutions for a number of problems in distributed control and sensor networks [25, 19, 10]. However, the works cited above assumed communications among nodes are either fixed or undirected.

Our paper is the first to demonstrate a working subgradient protocol in the setting of directed time-varying communications. We develop a broadcast-based protocol, termed the *subgradient-push*, which steers every node to an optimal value under a standard assumption of subgradient boundedness. The subgradient-push requires each node to know its out-degree at all times, but beyond this it needs no knowledge of the graph sequence or even of the number of agents to implement. Our results show that it converges at a rate of $O(\ln t/\sqrt{t})$, where the constant depends, among other factors, on the consensus speed of the corresponding directed graph sequence and a measure of the imbalance of influence among the nodes.

Our work is closest to the recent papers [28, 29, 6]. The papers [28, 29] proved the convergence of a subgradient algorithm in a directed but fixed topology; implementation of the protocol appears to require knowledge of the graph or of the number of agents. By contrast, our results work in time-varying networks and are fully distributed, requiring no knowledge of either the graph sequence or the number of agents. The paper [6] shows the convergence of a distributed optimization protocol in continuous time, also for directed but fixed graphs; moreover, an additional assumption is made in [6] that the graph is “balanced.”

All the prior work in distributed optimization, except for [28, 29] requires time-varying communications with some form of balancedness, often reflected in a requirement of having a sequence of doubly stochastic matrices that are commensurate with the sequence of underlying communication graphs. In contrast, our proposed method removes the need for the doubly stochastic matrices. The proposed distributed optimization model is motivated by applications that are characterized by time-varying directed communications such as those arising in a mobile sensor network communication where the links between nodes will come and go as nodes move in and out of line-of-sight or broadcast range of each other. Moreover, if different nodes are capable of broadcasting messages at different power levels, then communication links connecting the nodes will necessarily be unidirectional.

The remainder of this paper is organized as follows. We begin in Section 2 where we describe the problem of interest, outline the subgradient-push algorithm, and state the main convergence results. Section 3 is devoted to the proof of a key lemma, namely the convergence rate result for a perturbed version of the so-called push-sum protocol; this lemma is then used in the subsequent proofs of convergence and convergence rate for the subgradient-push in Section 4. Finally, some conclusions are offered in Section 5.

Notation: We will apply boldface to distinguish between the vectors in \mathbb{R}^d and scalars associated with different nodes. For example, the vector $\mathbf{x}_i(t)$ is in boldface to identify a vector for node i , while the scalar $y_i(t)$ is not - which identifies a scalar value for node i . Additionally, for a vector \mathbf{x}_i that has a subscript i identifying a node index, we will use $[\mathbf{x}]_j$

to denote its j 'th entry. The vectors such as $y(t) \in \mathbb{R}^n$ obtained by stacking scalar values $y_i(t)$ associated with the nodes is not bolded. For a matrix A , we will use $[A]_{ij}$ to denote the i, j 'th entry of A . The vectors are seen as column vectors unless otherwise explicitly stated. We use $\mathbf{1}$ to denote the vector of ones, and $\|y\|$ for the Euclidean norm of a vector y .

2 Problem, Algorithm and Main Results

We consider a network of n nodes whose goal is to minimize the function

$$F(\mathbf{z}) = \sum_{i=1}^n f_i(\mathbf{z})$$

where only node i knows the convex function $f_i(\mathbf{z}) : \mathbb{R}^d \rightarrow \mathbb{R}$. Under the assumption that the set of optimal solutions $Z^* = \arg \min_{\mathbf{z} \in \mathbb{R}^d} F(\mathbf{z})$ is nonempty, we would like to design a protocol in which all agents will maintain variables $\mathbf{z}_i(t)$ such that all the $\mathbf{z}_i(t)$ converge to the same point in Z^* .

We will assume that, at each time t , *node i can only send messages to its out-neighbors in some directed graph $G(t)$* . Naturally, the graph $G(t)$ will have vertex set $\{1, \dots, n\}$, and we will use $E(t)$ to denote its edge set. Also, naturally, the sequence $\{G(t)\}$ should possess some good long-term connectivity properties. A standard assumption, which we will be making, is that the sequence $\{G(t)\}$ is uniformly strongly connected (or, as it is sometimes called, B -strongly-connected), namely, that there exists some integer $B > 0$ (possibly unknown to the nodes) such that the graph with edge set

$$E_B(k) = \bigcup_{i=kB}^{(k+1)B-1} E(i)$$

is strongly connected for every $k \geq 0$. This is a typical assumption for many results in multi-agent control: it is considerably weaker than requiring each $G(t)$ be connected for it allows the edges necessary for connectivity to appear over a long time period and in arbitrary order; however, it is still strong enough to derive bounds on the speed of information propagation from one part of the network to another.

Finally, we introduce the notation $N_i^{\text{in}}(t)$ and $N_i^{\text{out}}(t)$ for the in- and out-neighborhoods of node i , respectively, at time t . We will allow these neighborhoods to include the node i itself¹; formally, we have

$$\begin{aligned} N_i^{\text{in}}(t) &= \{j \mid (j, i) \in E(t)\} \cup \{i\}, \\ N_i^{\text{out}}(t) &= \{j \mid (i, j) \in E(t)\} \cup \{i\}, \end{aligned}$$

and $d_i(t)$ for the out-degree of node i , i.e.,

$$d_i(t) = |N_i^{\text{out}}(t)|.$$

¹Alternatively, one may define these neighborhoods in a standard way of the graph theory, but require that each graph in the sequence $\{G(t)\}$ has a self-loop at every node.

Crucially, we will be assuming that *every node i knows its out-degree $d_i(t)$ at every time t .*

Our main result is a protocol which successfully accomplishes the task of distributed minimization of $F(\mathbf{z})$ under the assumptions we have laid out above. Our scheme is a combination of subgradient descent and the so-called *push-sum* protocol, recently studied in the papers [1, 4, 9]. We will refer to our protocol as the *subgradient-push* method.

2.1 The subgradient-push method

Every node i will maintain auxiliary vector variables $\mathbf{x}_i(t), \mathbf{w}_i(t)$ in \mathbb{R}^d , as well as an auxiliary scalar variable $y_i(t)$, initialized as $y_i(0) = 1$ for all i . These quantities will be updated by the nodes according to the rules,

$$\begin{aligned} \mathbf{w}_i(t+1) &= \sum_{j \in N_i^{\text{in}}(t)} \frac{\mathbf{x}_j(t)}{d_j(t)}, \\ y_i(t+1) &= \sum_{j \in N_i^{\text{in}}(t)} \frac{y_j(t)}{d_j(t)}, \\ \mathbf{z}_i(t+1) &= \frac{\mathbf{w}_i(t+1)}{y_i(t+1)}, \\ \mathbf{x}_i(t+1) &= \mathbf{w}_i(t+1) - \alpha(t+1)\mathbf{g}_i(t+1), \end{aligned} \tag{1}$$

where $\mathbf{g}_i(t+1)$ is a subgradient of the function f_i at $\mathbf{z}_i(t+1)$. The method is initiated with $\mathbf{w}_i(0) = \mathbf{z}_i(0) = \mathbf{1}$ and $y_i(0) = 1$ for all i . The stepsize $\alpha(t+1) > 0$ satisfies the following decay conditions

$$\sum_{t=1}^{\infty} \alpha(t) = \infty, \quad \sum_{t=1}^{\infty} \alpha^2(t) < \infty, \quad \alpha(t) \leq \alpha(s) \text{ for all } t > s \geq 0. \tag{2}$$

We note that the above equations have simple broadcast-based implementation: each node i broadcasts the quantities $\mathbf{x}_i(t)/d_i(t), y_i(t)/d_i(t)$ to all of the nodes in its out-neighborhood², which simply sum all the messages they receive to obtain $\mathbf{w}_i(t+1)$ and $y_i(t+1)$. The update equations for $\mathbf{z}_i(t+1), \mathbf{x}_i(t+1)$ can then be executed without any further communications between nodes during step t .

Without the subgradient term in the final equation, our protocol would be a version of the push-sum protocol [9] for average computation studied recently in [1, 4]. For intuition on the precise form of these equations, we refer the reader to these three papers; roughly speaking, the somewhat involved form of the updates is intended to ensure that every node receives an equal weighting after all the linear combinations and ratios have been taken. In this case, the vectors $\mathbf{z}_i(t+1)$ converge to some common point, i.e., a consensus is achieved. The inclusion of the subgradient terms in the updates of $\mathbf{x}_i(t+1)$ is intended to steer the consensus point towards the optimal set Z^* , while the push-sum updates steer the vectors $\mathbf{z}_i(t+1)$ towards each other. Our main results, which we describe in the next section, demonstrate that this scheme succeeds in steering all vectors $\mathbf{z}_i(t+1)$ towards the same point in the solution set Z^* .

²We note that we make use here of the assumption that node i knows its out-degree $d_i(t)$.

2.2 Our results

Our first theorem demonstrates the correctness of the subgradient-push method for an arbitrary stepsize $\alpha(t)$ satisfying Eq. (2); this holds under the assumptions we have laid out above, as well as an additional technical assumption on the boundedness of the subgradients.

Theorem 1 *Suppose that:*

- (a) *The graph sequence $\{G(t)\}$ is uniformly strongly connected with a self-loop at every node.*
- (b) *Each function $f_i(\mathbf{z})$ is convex and the set $Z^* = \arg \min_{\mathbf{z} \in \mathbb{R}^d} F(\mathbf{z})$ is nonempty.*
- (c) *The subgradients of each $f_i(\mathbf{z})$ are uniformly bounded, i.e., there exists $L_i < \infty$ such that*

$$\|\mathbf{g}_i\|_2 \leq L_i \quad \text{for all subgradients } \mathbf{g}_i \text{ of } f_i(\mathbf{z}) \text{ at all points } \mathbf{z} \in \mathbb{R}^d.$$

Then, the distributed subgradient-push method of Eq. (1) with the stepsize satisfying the conditions in Eq. (2) has the following property

$$\lim_{t \rightarrow \infty} \mathbf{z}_i(t) = \mathbf{z}^* \quad \text{for all } i \text{ and for some } \mathbf{z}^* \in Z^*.$$

Our second theorem makes explicit the rate at which the objective function converges to its optimal value. As standard with subgradient methods, we will make two tweaks in order to get a convergence rate result: (i) we take a stepsize which decays as $\alpha(t) = 1/\sqrt{t}$ (stepsizes which decay at faster rates usually produce inferior convergence rates), and (ii) each node i will maintain a convex combination of the values $\mathbf{z}_i(1), \mathbf{z}_i(2), \dots$ for which the convergence rate will be obtained. We then demonstrate that the subgradient-push converges at a rate of $O(\ln t/\sqrt{t})$; this is formally stated in the following theorem. The theorem makes use of the matrix $A(t)$ that captures the weights used in the construction of $\mathbf{w}_i(t+1)$ and $y_i(t+1)$ in Eq. (1), which are defined by

$$A_{ij}(t) = \begin{cases} 1/d_j(t) & \text{whenever } j \in N_i^{\text{in}}(t), \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Theorem 2 *Suppose all the assumptions of Theorem 1 hold and, additionally, $\alpha(t) = 1/\sqrt{t}$ for $t \geq 1$. Moreover, suppose that every node i maintains the variable $\tilde{\mathbf{z}}_i(t) \in \mathbb{R}^d$ initialized at time $t = 1$ to $\tilde{\mathbf{z}}_i(1) = \mathbf{z}_i(1)$ and updated as*

$$\tilde{\mathbf{z}}_i(t+1) = \frac{\alpha(t+1)\mathbf{z}_i(t+1) + S(t)\tilde{\mathbf{z}}_i(t)}{S(t+1)},$$

where $S(t) = \sum_{s=0}^{t-1} \alpha(s+1)$. Then, we have that for all $t \geq 1$, $i = 1, \dots, n$, and any $\mathbf{z}^ \in Z^*$,*

$$\begin{aligned} F(\tilde{\mathbf{z}}(t)) - F(\mathbf{z}^*) &\leq \frac{n \|\bar{\mathbf{x}}(0) - \mathbf{z}^*\|_1}{2\sqrt{t}} + \frac{n (\sum_{i=1}^n L_i)^2 (1 + \ln t)}{2 \cdot 4 \sqrt{t}} \\ &\quad + \frac{16}{\delta(1-\lambda)} \left(\sum_{i=1}^n L_i \right) \frac{\sum_{j=1}^n \|\mathbf{x}_j(0)\|_1}{\sqrt{t}} + \frac{16}{\delta(1-\lambda)} \left(\sum_{i=1}^n L_i^2 \right) \frac{(1 + \ln t)}{\sqrt{t}} \end{aligned}$$

where

$$\bar{\mathbf{x}}(0) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i(0),$$

and the scalars λ and δ are functions of the graph sequence $G(1), G(2), \dots$, which have the following properties:

(a) For any B -connected graph sequence with a self-loop at every node,

$$\begin{aligned} \delta &\geq \frac{1}{n^{nB}}, \\ \lambda &\leq \left(1 - \frac{1}{n^{nB}}\right)^{1/(nB)}. \end{aligned}$$

(b) If each of the graphs $G(t)$ is regular³, then

$$\begin{aligned} \delta &= 1 \\ \lambda &\leq \min \left\{ \left(1 - \frac{1}{4n^3}\right)^{1/B}, \max_{t \geq 1} \sqrt{\sigma_2(A(t))} \right\} \end{aligned}$$

where $A(t)$ is defined by Eq. (3) and $\sigma_2(A)$ is the second-largest singular value of a matrix A .

Several features of this theorem are expected: it is standard for a distributed subgradient method to converge at a rate of $O(\ln t / \sqrt{t})$ with the constant depending on the subgradient-norm upper bounds L_i , as well as on the initial conditions $\mathbf{x}_i(0)$ [23, 5]. Moreover, it is also standard for the analysis of these method to involve λ , which is a measure of the connectivity of the directed sequence $G(1), G(2), \dots$; namely, the closeness of λ to 1 measures the speed at which a consensus process on the graph sequence $\{G(t)\}$ converges.

However, our bounds also include the parameter δ , which, as we will later see, is a measure of the imbalance of influences among the nodes. Time-varying directed regular networks are uniform in influence and will have $\delta = 1$, so that δ will disappear from the bounds entirely; however, networks which are, in a sense to be specified, non-uniform will suffer a corresponding blow-up in the convergence time of the subgradient-push algorithm.

Moreover, we note that while the term $1/(\delta(1 - \lambda))$ appearing in our bounds is bounded only exponentially as n^{2nB} in the worst case, it need not be this large for every graph sequence; indeed, part (b) of Theorem 2 shows that for a class of time-varying directed graphs, $1/(\delta(1 - \lambda))$ scales polynomially in n . Our work therefore motivates the question of obtaining effective bounds on consensus speed and imbalance of the influence in sequences of directed graphs. Finally, we remark that previous research [17, 23, 5] has studied the case when the matrices $A(t)$ (defined in the statement of Theorem 2) are doubly stochastic; this occurs when the directed graph sequence $\{G(t)\}$ is regular, and in that case our polynomial bounds essentially match previously known results.

³The graph $G(t)$ is regular if there exists some $d(t)$ such that every out-degree and every in-degree of a node in $G(t)$ equals $d(t)$.

3 Perturbed Push-Sum Protocol

This section is dedicated to the analysis a perturbed version of the so-called push-sum protocol, originally introduced in the groundbreaking work [9] and recently analyzed in time-varying directed graphs in [1, 4]. The push-sum is a protocol for node interaction in directed topologies which allows nodes to compute averages and other aggregates in spite of the one-way nature of the communication links.

The original results of [9, 1, 4] demonstrate the convergence of the push-sum protocol. Here we will prove a generalization of this fact by showing that the protocol remains convergent even if the state of the nodes is perturbed at each step, as long as the size of the perturbations decays to zero. We will later use this result in the proof Theorems 1 and 2; because it has a self-contained interpretation and analysis, we sequester it to this section.

We begin with a statement of the perturbed push-sum update rule. Every node i maintains scalar variables $x_i(t), y_i(t), z_i(t), w_i(t)$ where we assume $y_i(0) = 1$ for all $i = 1, \dots, n$. These variables are updated as follows:

$$\begin{aligned} w_i(t+1) &= \sum_{j \in N_i^{\text{in}}(t)} \frac{x_j(t)}{d_j(t)}, \\ y_i(t+1) &= \sum_{j \in N_i^{\text{in}}(t)} \frac{y_j(t)}{d_j(t)}, \\ z_i(t+1) &= \frac{w_i(t+1)}{y_i(t+1)}, \\ x_i(t+1) &= w_i(t+1) + \epsilon_i(t+1), \end{aligned} \tag{4}$$

where $\epsilon_i(t)$ is a perturbation at every step, perhaps adversarially chosen. We assume that $N_i^{\text{in}}(t)$ is the in-neighborhood of node i in a directed graph $G(t)$ and $d_j(t)$ is the out-degree of node j , as previously defined in Section 2.

We note that without the perturbation term $\epsilon_i(t)$, the method in Eq. (4) reduces to the push-sum protocol. Moreover, our proposed subgradient-push method of Eq. (1) is simply Eq. (4) with a specific form for these perturbation vectors $\epsilon_i(t)$.

The precise form of the push-sum equations of Eq. (4) is a little involved. These dynamics were introduced for the purpose of average computation (in the case when all the perturbations $\epsilon_i(t)$ are zero) and have a simple motivating intuition. The push-sum is a variation of a consensus-like protocol wherein every node updates its values by taking linear combinations of the values of its neighbors; in such schemes, some nodes are bound to be more influential than others (meaning that other nodes end up placing higher coefficients on them), for example by virtue of being more centrally placed. The dynamics of the push-sum are designed around the ratio $z_i(t) = w_i(t)/y_i(t)$ in which these imbalances of influence are meant to be cancelled so that each $z_i(t)$ converges to $(1/n) \sum_{i=1}^n x_i(0)$. We refer the reader to [9, 4, 1] for more details.

We may rewrite the perturbed push-sum equations in more compact form. Using the definition of $A(t)$ from Eq. (3), the relations in Eq. (4) assume the following form:

$$w(t+1) = A(t)x(t),$$

$$\begin{aligned}
y(t+1) &= A(t)y(t), \\
z_i(t+1) &= \frac{w_i(t+1)}{y_i(t+1)}, \\
x(t+1) &= w(t+1) + \epsilon(t+1),
\end{aligned} \tag{5}$$

where $\epsilon(t) = (\epsilon_1(t), \dots, \epsilon_n(t))'$. Observe that each of the matrices $A(t)$ is column-stochastic but not necessarily row-stochastic.

We will be concerned here with demonstrating a convergence result and a convergence rate for the updates given in Eq. (4), or equivalently, in Eq. (5). Specifically, the bulk of this section is dedicated to proving the following lemma.

Lemma 3 *Consider the sequences $\{z_i(t)\}$, $i = 1, \dots, n$, generated by the method in Eq. (4). Assuming that the graph sequence $\{G(t)\}$ is uniformly strongly connected, the following statements hold:*

(a) *There exists some $\delta > 0$ and $\lambda \in (0, 1)$ such that for all $t \geq 1$ we have*

$$\left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| \leq \frac{8}{\delta} \left(\lambda^t \|x(0)\|_1 + \sum_{s=1}^t \lambda^{t-s} \|\epsilon(s)\|_1 \right),$$

Moreover, we may choose δ, λ satisfying

$$\delta \geq \frac{1}{n^{nB}}, \quad \lambda \leq \left(1 - \frac{1}{n^{nB}} \right)^{1/B}.$$

If in addition each of the matrices $A(t)$ is doubly stochastic, then

$$\delta = 1, \quad \lambda \leq \left\{ \left(1 - \frac{1}{4n^3} \right)^{1/B}, \max_{t \geq 1} \sqrt{\sigma_2(A(t))} \right\}.$$

(b) *If $\lim_{t \rightarrow 0} \epsilon_i(t) = 0$ for all $i = 1, \dots, n$, then*

$$\lim_{t \rightarrow 0} \left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| = 0.$$

(c) *If $\{\alpha(t)\}$ is a non-increasing positive scalar sequence such that $\sum_{t=1}^{\infty} \alpha(t) |\epsilon_i(t)| < \infty$ for all i , then*

$$\sum_{t=0}^{\infty} \alpha(t+1) \left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| < \infty \quad \text{for all } i = 1, \dots, n.$$

For part (b) of Lemma 3, observe that each of the matrices $A(t)$ is doubly stochastic if each of the graphs $G(t)$ is regular. Furthermore, we observe that if $\epsilon_i(t) = 0$, this lemma implies that the push-sum method converges at a geometric rate; moreover, it is easy to see that $\mathbf{1}'x(t)/n = \mathbf{1}'x(0)/n$ and therefore $z_i(t) \rightarrow \mathbf{1}'x(0)/n$, so that the push-sum protocol successfully computes the average. In the more general case when the perturbations are nonzero, the lemma states that if these perturbations decay to zero, then the push-sum method still converges. Of course, it will no longer be true in this case that the convergence is necessarily on the average of the initial values.

We will prove a series of auxiliary lemmas before beginning the proof of Lemma 3. We first remark that the matrices $A(t)$ have a special structure that allows us to efficiently analyze their products. Specifically, we have the following properties of the matrices $A(t)$ (see [2, 7, 14, 18, 30] for proofs of this and similar statements).

Lemma 4 *Suppose that the graph sequence $\{G(t)\}$ is uniformly strongly-connected. Then, the following statements are true:*

- (a) *For every $s \geq 0$, the limit $\lim_{t \rightarrow \infty} A'(t)A'(t-1) \cdots A'(s+1)A'(s)$ exists. In particular, the limiting matrix is a rank-one stochastic matrix, i.e., there is a stochastic vector $\phi(s)$ such that*

$$\lim_{t \rightarrow \infty} A'(t)A'(t-1) \cdots A'(s+1)A'(s) = \mathbf{1}\phi'(s) \quad \text{for all } s \geq 0.$$

- (b) *The convergence rate is geometric*

$$|[A'(t)A'(t-1) \cdots A'(s+1)A'(s)]_{ij} - \phi_j(s)| \leq C\lambda^{t-s} \quad \text{for all } i, j = 1, \dots, n,$$

for some C and $\lambda \in (0, 1)$.

There are also known bounds on the parameters C, λ from this lemma which upper bound how large C is and how far away λ is from 1. Moreover, these bounds improve if the sequence $\{G(t)\}$ has some nice properties. The following lemma is a formal statement to this effect.

Lemma 5 *We have:*

- (a) *For any B -strongly connected sequence of graphs, we may choose*

$$C = 2, \quad \lambda = \left(1 - \frac{1}{n^{nB}}\right)^{1/B}$$

in the statement of Lemma 4.

- (b) *If in addition every graph is regular, we may choose*

$$C = \sqrt{2}, \quad \lambda = \min \left\{ \left(1 - \frac{1}{4n^3}\right)^{1/B}, \max_{t \geq 1} \sqrt{\sigma_2(A(t))} \right\}.$$

Proof. From [2, 7, 30], under the assumption of B -strong connectivity and our definition of neighborhoods, we have that if

$$x(t) = A'(t-1) \cdots A'(s)x(s),$$

then

$$\max_{i=1,\dots,n} x_i(t) - \min_{i=1,\dots,n} x_i(t) \leq \left(1 - \frac{1}{n^{nB}}\right)^{\lfloor (t-s)/(nB) \rfloor} \left(\max_{i=1,\dots,n} x_i(s) - \min_{i=1,\dots,n} x_i(s) \right).$$

We note that this implies

$$\max_{i=1,\dots,n} x_i(t) - \min_{i=1,\dots,n} x_i(t) \leq 2 \left(\left(1 - \frac{1}{n^{nB}}\right)^{1/(nB)} \right)^{t-s} \left(\max_{i=1,\dots,n} x_i(s) - \min_{i=1,\dots,n} x_i(s) \right).$$

This holds for every $x(s)$. By choosing $x(s)$ to be each of the n basis vectors, we see that for every $j = 1, \dots, n$,

$$\max_i [A'(t) \cdots A'(s)]_{ij} - \min_i [A'(t) \cdots A'(s)]_{ij} \leq 2 \left(\left(1 - \frac{1}{n^{nB}}\right)^{1/(nB)} \right)^{t-s}.$$

Since $\phi_j(s)$ is a convex combination of the n numbers $[A'(t) \cdots A'(s)]_{ij}$, $i = 1, \dots, n$, we have proven part (a).

As for the second statement, when each of the matrices $A(t)$ is doubly stochastic, the results of [16] imply that if

$$x(t) = A'(t-1) \cdots A'(s)x(s),$$

then

$$\sum_{i=1}^n (x_i(t) - \bar{x})^2 \leq \left(1 - \frac{1}{2n^3}\right)^{\lfloor (t-s)/B \rfloor} \sum_{i=1}^n (x_i(s) - \bar{x})^2,$$

where \bar{x} is the average entry of $x(s)$. Similarly, we write this as

$$\sum_{i=1}^n (x_i(t) - \bar{x})^2 \leq 2 \left(\left(1 - \frac{1}{2n^3}\right)^{1/B} \right)^{t-s} \sum_{i=1}^n (x_i(s) - \bar{x})^2.$$

Moreover, plugging in each basis vector, we obtain that for each j ,

$$\max_{i=1,\dots,n} [A'(t) \cdots A'(s)]_{ij} - \frac{1}{n} \leq \sqrt{2} \left(\left(\sqrt{1 - \frac{1}{2n^3}} \right)^{1/B} \right)^{t-s}.$$

Since $\sqrt{1 - \beta/2} \leq 1 - \beta/4$ for all $\beta \in (0, 1)$, this implies that we may choose $C = \sqrt{2}$, $\lambda = (1 - 1/(4n^3))^{1/B}$. The same line of argument shows that we may choose $C = 1$ and $\lambda = \max_{t \geq 1} \sqrt{\sigma_2(A(t))}$. ■

The next lemma provides a bound on how small the entries of the products $A'(t) \cdots A'(1)$ can get. We will use these bounds later in the proof of Lemma 4.

Lemma 6 *Given a graph sequence $\{G(t)\}$, define*

$$\delta = \inf_{t=1,2,\dots} \min_{i=1,\dots,n} [\mathbf{1}' A'(t) \cdots A'(1)]_i.$$

If $G(t)$ is B -strongly connected, then

$$\delta \geq \frac{1}{n^{nB}}.$$

If each $G(t)$ is regular, then

$$\delta = 1.$$

Proof. By the definition of matrices $A(t)$ in Eq. (3), we have that for all $t \geq 1$,

$$[A'(t+1) \cdots A'(1)]_{ii} \geq \frac{1}{n} [A'(t) \cdots A'(1)]_{ii},$$

and again due to the presence of self-loops, $[A'(1)]_{ii} \geq 1/n$ for all i . Thus, we certainly have that $[\mathbf{1}' A'(t) \cdots A'(1)]_i \geq 1/n^{nB}$ for all i and all t in the range $1 \leq t \leq n^{nB}$. However, it was shown in [7, 30] that for $t > (n-1)B$, every entry of $A'(t) \cdots A'(1)$ is positive and has value at least $1/n^{nB}$. Since $n^{nB} > (n-1)B$, this proves the bound $\delta \geq 1/n^{nB}$. The final claim that $\delta = 1$ for a sequence of regular graphs is trivial. ■

Remark 7 *Observe that as an immediate consequence of the definition of δ , we have $\phi_j(s) \geq \delta/n$ for all $j = 1, \dots, n$.*

By taking transposes and applying the previous lemmas, we immediately get the following result on the products $A(t) \cdots A(s)$; for convenience, let us adopt the notation of referring to these products as $A(t : s)$.

Corollary 8 *Suppose that the graph sequence $\{G(t)\}$ is B -strongly-connected. Then:*

- (a) *There is a sequence of stochastic vectors $\phi(t)$ such that the matrix difference $A(t : s) - \phi(t)\mathbf{1}'$ for $t \geq s$ decays geometrically, i.e.,*

$$|[A(t : s)]_{ij} - \phi_i(t)| \leq C\lambda^{t-s} \quad \text{for all } i, j = 1, \dots, n,$$

where we can always choose

$$C = 4, \lambda = \left(1 - 1/n^{nB}\right)^{1/B}.$$

If in addition each $A(t)$ is doubly stochastic, we may choose

$$C = 2\sqrt{2}, \lambda = \left(1 - 1/(4n^3)\right)^{1/B}$$

or

$$C = 2, \lambda = \max_t \sqrt{\sigma_2(A(t))},$$

whenever the last quantity is below 1.

(b) The quantity

$$\delta = \inf_{t=1,2,\dots} \min_{i=1,\dots,n} [\mathbf{1}A'(t) \cdots A'(1)]_i.$$

satisfies

$$\delta \geq \frac{1}{n^{nB}}.$$

Moreover, if the graphs $G(t)$ are regular, we have $\delta = 1$. Furthermore, we have

$$\phi_j(t) \geq \frac{\delta}{n} \quad \text{for all times } t.$$

Proof. The lemma follows by taking transposes and applying Lemmas 4, 5, 6 as well as Remark 7. The only thing that needs to be proved is that these lemmas can be applied, which is routine, with the exception of the issue of B -connectivity, which requires further elaboration.

When we transpose $A(t : s)$ to get the product $A'(s) \cdots A'(t-1)A'(t)$, we have reversed the order in which the matrices appear; and, moreover, by taking the transposes of each matrix, we have effectively reversed the direction of every edge in each $G(t)$. We must thus argue that when we take an initial segment of B -connected graph sequence and reverse the order of the graphs as well as the direction of each edge, we still have an initial segment of a B -connected sequence. Unfortunately, this is not true; however, it is easy to see that it is true after we throw out at most B graphs from the start of the sequence. The bound of Lemma 4 part (b) thus applies with $t-s$ replaced by $t-s-B$, which we take care of by doubling the constants C instead. ■

We remark that δ as defined in Lemma 6 may be thought of as a measure of imbalance of the influence among the nodes. Indeed, δ is defined as the best lower bound on the row sums of the matrices $A(t : s)$. In the case when each of the graphs $G(t)$ is regular, the matrices $A(t)$ will be doubly stochastic and we will have $\delta = 1$ as previously remarked (i.e., no imbalance of influence). By contrast, when $\delta \approx 0$, there is a node i such that the i 'th row in some $A(t : s)$ will have entries which are all nearly zero; in short, it is almost as if node i has no in-neighbors at all and it is uninfluenced by what occurs in the rest of the network. In cases intermediate between these two extremes, δ reflects the existence of a node whose influence on the rest of the network, measured by summing up all the weights placed on it in $A(t : s)$, is small.

We proceed with our sequence of intermediate lemmas for the proof of Lemma 3. We will now need some auxiliary results on the convolution of two scalar sequences, as in the following statement.

Lemma 9 ([23] Lemma 3.1) *Let $\{\gamma_k\}$ be a scalar sequence.*

(a) *If $\lim_{k \rightarrow \infty} \gamma_k = \gamma$ and $0 < \beta < 1$, then $\lim_{k \rightarrow \infty} \sum_{\ell=0}^k \beta^{k-\ell} \gamma_\ell = \frac{\gamma}{1-\beta}$.*

(b) *If $\gamma_k \geq 0$ for all k , $\sum_{k=0}^{\infty} \gamma_k < \infty$ and $0 < \beta < 1$, then $\sum_{k=0}^{\infty} \left(\sum_{\ell=0}^k \beta^{k-\ell} \gamma_\ell \right) < \infty$.*

(c) *If $\limsup_{k \rightarrow \infty} \gamma_k = \gamma$ and $\{\zeta_k\}$ is a positive scalar sequence with $\sum_{k=0}^{\infty} \zeta_k = \infty$, then $\limsup_{k \rightarrow \infty} \frac{\sum_{\ell=0}^k \gamma_\ell \zeta_\ell}{\sum_{\ell=0}^k \zeta_\ell} \leq \gamma$. In addition, if $\liminf_{k \rightarrow \infty} \gamma_k = \gamma$, then $\lim_{k \rightarrow \infty} \frac{\sum_{\ell=0}^k \gamma_\ell \zeta_\ell}{\sum_{\ell=0}^k \zeta_\ell} = \gamma$.*

With these pieces in places, we can now proceed to the proof of Lemma 3. Our argument will rely on Corollary 8 on the products $A(t : s)$ and the just-stated Lemma 9 on sequence convolutions.

Proof of Lemma 3. (a) By inspecting Eq. (5) it is easy to see that

$$x(t+1) = A(t : 0)x(0) + \sum_{s=1}^t A(t : s)\epsilon(s) + \epsilon(t+1). \quad (6)$$

which implies

$$A(t+1)x(t+1) = A(t+1 : 0)x(0) + \sum_{s=1}^{t+1} A(t+1 : s)\epsilon(s). \quad (7)$$

Moreover, since each $A(t)$ is column-stochastic, we have that $\mathbf{1}'A(t) = \mathbf{1}'$ and Eq. (6) also implies that

$$\mathbf{1}'x(t+1) = \mathbf{1}'x(0) + \sum_{s=1}^{t+1} \mathbf{1}'\epsilon(s). \quad (8)$$

Now Eq. (7) and Eq. (8) give us that

$$\begin{aligned} A(t+1)x(t+1) - \phi(t+1)\mathbf{1}'x(t+1) &= (A(t+1 : 0) - \phi(t+1)\mathbf{1}')x(0) \\ &\quad + \sum_{s=1}^{t+1} (A(t+1 : s) - \phi(t+1)\mathbf{1}')\epsilon(s). \end{aligned} \quad (9)$$

However, according to Corollary 8, if we define $D(t, s)$ to be

$$D(t, s) = A(t : s) - \phi(t)\mathbf{1}'$$

then we have the entry-wise decay bound

$$|[D(t, s)]_{ij}| \leq C\lambda^{t-s} \quad \text{for all } i, j \text{ and } t \geq 0, \quad (10)$$

for some constants $C > 0$ and $\lambda \in (0, 1)$. Moreover, the constants C and λ have all properties listed in Corollary 8.

Therefore, from relation (9) it follows for $t \geq 0$,

$$A(t+1)x(t+1) = \phi(t+1)\mathbf{1}'x(t+1) + D(t+1, 0)x(0) + \sum_{s=1}^{t+1} D(t+1, s)\epsilon(s).$$

Thus, for $t \geq 1$ we have

$$w(t+1) = A(t)x(t) = \phi(t)\mathbf{1}'x(t) + D(t, 0)x(0) + \sum_{s=1}^t D(t, s)\epsilon(s). \quad (11)$$

We may derive a similar expression for $y(t+1)$:

$$y(t+1) = A(t : 0)y(0) = \phi(t)\mathbf{1}'y(0) + D(t, 0)y(0) = \phi(t)n + D(t : 0)\mathbf{1}, \quad (12)$$

From (11) and (12) we obtain for every $t \geq 1$ and all i ,

$$z_i(t+1) = \frac{w_i(t+1)}{y_i(t+1)} = \frac{\phi_i(t) \mathbf{1}'x(t) + [D(t:0)x(0)]_i + \sum_{s=1}^t [D(t:s)\epsilon(s)]_i}{\phi_i(t) n + [D(t:0)\mathbf{1}]_i}.$$

Therefore,

$$\begin{aligned} z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} &= \frac{\phi_i(t) \mathbf{1}'x(t) + [D(t:0)x(0)]_i + \sum_{s=1}^t [D(t:s)\epsilon(s)]_i}{\phi_i(t) n + [D(t:0)\mathbf{1}]_i} - \frac{\mathbf{1}'x(t)}{n} \\ &= \frac{n ([D(t:0)x(0)]_i + \sum_{s=1}^t [D(t:s)\epsilon(s)]_i) - \mathbf{1}'x(t)[D(t:0)\mathbf{1}]_i}{n (\phi_i(t) n + [D(t:0)\mathbf{1}]_i)}. \end{aligned}$$

Observe that the denominator of the above fraction is n times the i 'th row sum of $A(t:0)$. By definition of δ , this row sum is at least δ , and consequently

$$n (\phi_i(t) n + [D(t:0)\mathbf{1}]_i) \geq n\delta$$

Thus,

$$\begin{aligned} \left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| &\leq \frac{|n ([D(t:0)x(0)]_i + \sum_{s=1}^t [D(t:s)\epsilon(s)]_i)| + |\mathbf{1}'x(t)[D(t:0)\mathbf{1}]_i|}{n |\phi_i(t) n + [D(t:0)\mathbf{1}]_i|} \\ &\leq \frac{1}{n\delta} \left(n \left(\max_j |[D(t:0)]_{ij}| \right) \|x(0)\|_1 + n \sum_{s=1}^t \left(\max_j |[D(t:s)]_{ij}| \right) \|\epsilon(s)\|_1 \right. \\ &\quad \left. + |\mathbf{1}'x(t)| \left(\max_j |[D(t:0)]_{ij}| \right) n \right). \end{aligned}$$

Factoring an n out and using estimates for $|[D(t:s)]_{ij}|$ as given in (10), we find

$$\left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| \leq \frac{C}{\delta} \left(\lambda^t \|x(0)\|_1 + \sum_{s=1}^t \lambda^{t-s} \|\epsilon(s)\|_1 + |\mathbf{1}'x(t)| \lambda^t \right).$$

Now we look at the term $\mathbf{1}'x(t)$. From Eq. (8), we have

$$|\mathbf{1}'x(t)| \leq \|x(0)\|_1 + \sum_{s=1}^t \|\epsilon(s)\|_1.$$

Going back to our estimate,

$$\begin{aligned} \left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| &\leq \frac{C}{\delta} \left(\lambda^t \|x(0)\|_1 + \sum_{s=1}^t \lambda^{t-s} \|\epsilon(s)\|_1 + \lambda^t \left(\|x(0)\|_1 + \sum_{s=1}^t \|\epsilon(s)\|_1 \right) \right) \\ &= \frac{C}{\delta} \left(2\lambda^t \|x(0)\|_1 + 2 \sum_{s=1}^t \lambda^{t-s} \|\epsilon(s)\|_1 \right). \end{aligned}$$

Since we were able to choose $C \leq 4$ in all the cases considered in Lemma 8, we may choose $C = 4$ to obtain a proof of part (a).

(b) By letting $t \rightarrow \infty$ in the preceding relation, since $\lambda \in (0, 1)$, we find that

$$\lim_{t \rightarrow \infty} \left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| \leq \lim_{t \rightarrow \infty} \sum_{s=1}^t \lambda^{t-s} \|\epsilon(s)\|_1.$$

When $\epsilon_i(t) \rightarrow 0$ for all i , then $\|\epsilon(t)\|_1 \rightarrow 0$ and, by Lemma 9(a), we conclude that

$$\lim_{t \rightarrow \infty} \sum_{s=1}^t \lambda^{t-s} \|\epsilon(s)\|_1 = 0,$$

and the result follows from the preceding two relations.

(c) Since $\{\alpha(t)\}$ is positive and non-increasing sequence, we have $\alpha(t+1) \leq \alpha(0)$ and $\alpha(t+1) \leq \alpha(s)$ for all $s \leq t$. Using these relations, we obtain

$$\alpha(t+1) \left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| \leq \frac{2C}{\delta} \left(\alpha(0) \lambda^t \|x(0)\|_1 + \sum_{s=1}^t \lambda^{t-s} \alpha(s) \|\epsilon(s)\|_1 \right). \quad (13)$$

Since $\lambda \in (0, 1)$, the sum $\sum_{t=1}^{\infty} \lambda^t$ is finite, and by Lemma 9(b) the sum $\sum_{t=1}^{\infty} \sum_{\ell=1}^t \lambda^{t-\ell} \alpha(\ell) \|\epsilon(\ell)\|_1$ is finite. Therefore

$$\sum_{t=1}^{\infty} \alpha(t+1) \left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| < \infty.$$

■

Lemma 3, which we have just proved, is the central result of this section; it states that each of the sequences $z_i(t+1)$ tracks the average $\bar{x}(t) = \mathbf{1}'x(t)/n$ increasingly well as time goes on. We will later require a corollary of this lemma: we will need the fact that a weighted average of the $z_i(t+1)$ tracks a weighted averages of $\bar{x}(t)$. The next corollary gives a precise statement of this.

Corollary 10 *Suppose all the assumptions of Lemma 3 are satisfied, and moreover $\alpha(t) = 1/\sqrt{t}$ and the size of the perturbations $\epsilon_i(t)$ is bounded as $\|\epsilon_i(t)\|_1 \leq D/\sqrt{t}$. Defining*

$$\bar{x}(t) = \frac{1}{n} \sum_{i=1}^n x_i(t),$$

we then have that for every $i = 1, \dots, n$,

$$\frac{1}{\sum_{k=0}^t \alpha(k+1)} \sum_{k=0}^t \alpha(k+1) |z_i(k+1) - \bar{x}(k)| \leq 4 \frac{\|x(0)\|_1 + D(1 + \ln t)}{\delta(1 - \lambda)(\sqrt{t+2} - 1)}.$$

Proof. Indeed, on the one hand

$$\sum_{k=0}^t \alpha(k+1) \geq \int_0^{t+1} \frac{1}{\sqrt{u+1}} du = 2 \left(\sqrt{t+2} - 1 \right). \quad (14)$$

On the other hand using Lemma 3,

$$\begin{aligned}
\sum_{k=0}^t \alpha(k+1) |\bar{x}(k) - z_i(k+1)| &\leq \frac{8}{\delta} \sum_{k=0}^t \frac{\lambda^k}{\sqrt{k+1}} \|x(0)\|_1 + \frac{8}{\delta} \sum_{k=1}^t \alpha(k+1) \sum_{s=1}^k \lambda^{k-s} \|\epsilon(s)\|_1 \\
&\leq \frac{8}{\delta} \frac{1}{1-\lambda} \|x(0)\|_1 + \frac{8D}{\delta} \sum_{k=1}^t \sum_{s=1}^k \frac{\lambda^{k-s}}{s} \\
&\leq \frac{8}{\delta(1-\lambda)} \|x(0)\|_1 + \frac{8D}{\delta} \frac{(1+\ln t)}{1-\lambda}.
\end{aligned} \tag{15}$$

■

We conclude this section by remarking that Lemma 3 and Corollary 10 hold even if $x_i(t)$ (and, by extension, $z_i(t)$) is a d -dimensional vector, by applying the results to each coordinate component of the space.

4 Convergence Results for Subgradient-Push Method

We turn now to the proofs of our main results, Theorems 1 and 2. Our arguments will crucially rely on the convergence results for the perturbed push-sum method we have established in the previous section.

We give a brief, informal summary of the main ideas behind our argument. The convergence result for the perturbed push-sum method of the previous section implies that, under the appropriate assumptions, the entries of $\mathbf{z}_i(t)$ get close to each other over time, and consequently $\mathbf{z}_i(t)$ approaches a multiple of the all-ones vector. Thus every node takes a subgradient of its own function f_i at nearly the same point; over time, these subgradients are averaged by the push-sum-like updates of our method, and the subgradient push approximates the ordinary subgradient algorithm applied to the average function $\frac{1}{n} \sum_{j=1}^n f_j$.

We now begin the formal process of proving Theorems 1 and 2. Our first step will be to establish two lemmas pertaining to the convergence of a scalar sequence. The first is a deterministic counterpart of the well-known supermartingale convergence result ([26]; see also [21], Lemma 11, Chapter 2.2).

Lemma 11 *Let $\{v_t\}$ be a non-negative scalar sequence such that*

$$v_{t+1} \leq (1 + b_t)v_t - u_t + c_t \quad \text{for all } t \geq 0,$$

where $b_t \geq 0$, $u_t \geq 0$ and $c_t \geq 0$ for all k with $\sum_{t=0}^{\infty} b_t < \infty$, and $\sum_{t=0}^{\infty} c_t < \infty$. Then the sequence $\{v_t\}$ converges to some $v \geq 0$ and $\sum_{t=0}^{\infty} u_t < \infty$.

We can use this lemma to derive the convergence of a sequence satisfying a subgradient-like recursion, as in the following lemma.

Lemma 12 *Consider a convex minimization problem $\min_{x \in \mathbb{R}^m} f(x)$, where $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is a convex function and $X \subseteq \mathbb{R}^m$ is a convex closed set. Assume that the solution set X^* of the problem is nonempty. Let $\{x_t\}$ be a sequence such that for all $x \in X$,*

$$\|x_{t+1} - x\|^2 \leq (1 + b_t)\|x_t - x\|^2 - \alpha_t (f(x_t) - f(x)) + c_t \quad \text{for all } t \geq 0,$$

where $b_t \geq 0$, $\alpha_t \geq 0$ and $c_t \geq 0$ for all k with $\sum_{t=0}^{\infty} b_t < \infty$, $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} c_t < \infty$. Then, the sequence $\{x_t\}$ converges to some solution x^* of the problem.

Proof. By letting $x = x^*$ for arbitrary $x^* \in X^*$, we obtain

$$\|x_{t+1} - x^*\|^2 \leq (1 + b_t)\|x_t - x^*\|^2 - \alpha_t(f(x_t) - f(x^*)) + c_t \quad \text{for all } t \geq 0.$$

Thus, all the conditions of Lemma 11 are satisfied, and by this lemma we obtain the following statements:

$$\text{the sequence } \{\|x_t - x^*\|^2\} \text{ is convergent for every } x^* \in X^*, \quad (16)$$

$$\sum_{t=0}^{\infty} \alpha_t(f(x_t) - f^*) < \infty, \quad (17)$$

where $f^* = \min_{x \in X} f(x)$. Since $\sum_{t=0}^{\infty} \alpha_t = \infty$, it follows from (17) that

$$\liminf_{t \rightarrow \infty} f(x_t) = f^*.$$

Let $\{x_{t_\ell}\}$ be a subsequence of $\{x_t\}$ such that

$$\lim_{\ell \rightarrow \infty} f(x_{t_\ell}) = \liminf_{t \rightarrow \infty} f(x_t) = f^*. \quad (18)$$

Now Eq. (16) implies that the sequence x_t takes values in a compact set, thus we can assume without loss of generality that $\{x_{t_\ell}\}$ is converging to some \tilde{x} (for otherwise, we can in turn select a convergent subsequence of $\{x_{t_\ell}\}$). The function f is convex over \mathbb{R}^m , so it is continuous. Therefore,

$$\lim_{\ell \rightarrow \infty} f(x_{t_\ell}) = f(\tilde{x}),$$

which by (18) implies that $\tilde{x} \in X^*$. By letting $x^* = \tilde{x}$ in (16) we obtain that the whole sequence $\{x_t\}$ must converge to \tilde{x} . ■

A key step in the proofs of Theorems 1 and 2 will be obtained by applying the previous lemma to the average-value process

$$\bar{\mathbf{x}}(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i(t),$$

where $\mathbf{x}_i(t)$ is the sequence generated by the subgradient-push method. We will need to argue that $\bar{\mathbf{x}}(t)$ satisfies the assumptions of the previous lemma, for which the following statement will be instrumental.

Lemma 13 *Under the same assumptions as in Theorem 1, we have for all $\mathbf{v} \in \mathbb{R}^d$ and $t \geq 0$:*

$$\begin{aligned} \|\bar{\mathbf{x}}(t+1) - \mathbf{v}\|^2 &\leq \|\bar{\mathbf{x}}(t) - \mathbf{v}\|^2 - \frac{2\alpha(t+1)}{n} (F(\bar{\mathbf{x}}(t)) - F(\mathbf{v})) \\ &\quad + \frac{4\alpha(t+1)}{n} \sum_{i=1}^n L_i \|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| + \alpha^2(t+1) \frac{(\mathbf{1}' L_F)^2}{n^2}. \end{aligned}$$

Proof. Let us define $\tilde{x}_\ell(t)$ to be the vector in \mathbb{R}^n which stacks up the ℓ 'th entries of all the vectors $\mathbf{x}_i(t)$: formally, we define $\tilde{x}_\ell(t)$ to be the vector whose j 'th entry is the ℓ 'th entry of $\mathbf{x}_j(t)$. Similarly, we define $\tilde{g}_\ell(t)$ to be the vector stacking up the ℓ 'th entries of the vectors $\mathbf{g}_i(t)$: the j 'th entry of $\tilde{g}_\ell(t)$ is the ℓ 'th entry of $\mathbf{g}_j(t)$.

It is easy to see from the definition of the subgradient-push in Eq. (1) that

$$\tilde{x}_\ell(t+1) = A(t)\tilde{x}_\ell(t) - \alpha(t+1)\tilde{g}_\ell(t+1) \quad \text{for } \ell = 1, \dots, d.$$

Since $A(t)$ is a column-stochastic matrix, this implies for all $\ell = 1, \dots, d$,

$$\frac{1}{n} \sum_{j=1}^n [\tilde{x}_\ell(t+1)]_j = \frac{1}{n} \sum_{j=1}^n [\tilde{x}_\ell(t)]_j - \frac{\alpha(t+1)}{n} \sum_{j=1}^n [\tilde{g}_\ell(t+1)]_j.$$

Since the ℓ 'th entry of $\bar{\mathbf{x}}(t+1)$ is exactly the left-hand side above, we can conclude that

$$\bar{\mathbf{x}}(t+1) = \bar{\mathbf{x}}(t) - \frac{\alpha(t+1)}{n} \sum_{j=1}^n \mathbf{g}_j(t+1). \quad (19)$$

Now let $\mathbf{v} \in \mathbb{R}^d$ be an arbitrary vector. From relation (19) it follows that for all $t \geq 0$,

$$\|\bar{\mathbf{x}}(t+1) - \mathbf{v}\|^2 = \|\bar{\mathbf{x}}(t) - \mathbf{v}\|^2 - \frac{2\alpha(t+1)}{n} \sum_{i=1}^n \mathbf{g}'_i(t+1)(\bar{\mathbf{x}}(t) - \mathbf{v}) + \frac{\alpha^2(t+1)}{n^2} \left\| \sum_{i=1}^n \mathbf{g}_i(t+1) \right\|^2.$$

Since the subgradient norms of each f_i are uniformly bounded by L_i , it further follows that for all $t \geq 0$,

$$\|\bar{\mathbf{x}}(t+1) - \mathbf{v}\|^2 \leq \|\bar{\mathbf{x}}(t) - \mathbf{v}\|^2 - \frac{2\alpha(t+1)}{n} \sum_{i=1}^n \mathbf{g}'_i(t+1)(\bar{\mathbf{x}}(t) - \mathbf{v}) + \alpha^2(t+1) \frac{(\mathbf{1}' L_F)^2}{n^2}, \quad (20)$$

where $L_F = (L_1, \dots, L_n)'$.

We next consider the cross-term $\sum_{i=1}^n \mathbf{g}'_i(t+1)(\bar{\mathbf{x}}(t) - \mathbf{v})$ in (20). For this term, we write

$$\sum_{i=1}^n \mathbf{g}'_i(t+1)(\bar{\mathbf{x}}(t) - \mathbf{v}) = \sum_{i=1}^n \mathbf{g}'_i(t+1) ((\bar{\mathbf{x}}(t) - \mathbf{z}_i(t+1)) + (\mathbf{z}_i(t+1) - \mathbf{v})). \quad (21)$$

Using the subgradient boundedness and Cauchy-Schwarz, we can lower bound the first term $\mathbf{g}'_i(t+1)(\bar{\mathbf{x}}(t) - \mathbf{z}_i(t+1))$ as

$$\mathbf{g}'_i(t+1)(\bar{\mathbf{x}}(t) - \mathbf{z}_i(t+1)) \geq -L_i \|\bar{\mathbf{x}}(t) - \mathbf{z}_i(t+1)\|. \quad (22)$$

As for the second term $\mathbf{g}'_i(t+1)(\mathbf{z}_i(t+1) - \mathbf{v})$, we can use the fact that $\mathbf{g}'_i(t+1)$ is the subgradient of $f_i(\theta)$ at $\theta = \mathbf{z}_i(t+1)$ to obtain:

$$\mathbf{g}_i(t+1)(\mathbf{z}_i(t+1) - \mathbf{v}) \geq f_i(\mathbf{z}_i(t+1)) - f_i(\mathbf{v}),$$

from which, by adding and subtracting $\bar{\mathbf{x}}(t)$ and using the Lipschitz continuity of f_i (implied by the subgradient boundedness), we further obtain

$$\mathbf{g}_i(t+1)(\mathbf{z}_i(t+1) - \mathbf{v}) \geq f_i(\mathbf{z}_i(t+1)) - f_i(\bar{\mathbf{x}}(t)) + f_i(\bar{\mathbf{x}}(t)) - f_i(\mathbf{v})$$

$$\geq -L_i \|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| + f_i(\bar{\mathbf{x}}(t)) - f_i(\mathbf{v}). \quad (23)$$

By substituting the estimates obtained in (22)–(23) back in relation (21), and using $F(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x})$ we obtain

$$\sum_{i=1}^n \mathbf{g}'_i(t+1)(\bar{\mathbf{x}}(t) - \mathbf{v}) \geq F(\bar{\mathbf{x}}(t)) - F(\mathbf{v}) - 2 \sum_{i=1}^n L_i \|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\|. \quad (24)$$

Now, we substitute estimate (24) into relation (20) and obtain for any $\mathbf{v} \in \mathbb{R}^d$ and all $t \geq 0$,

$$\begin{aligned} \|\bar{\mathbf{x}}(t+1) - \mathbf{v}\|^2 &\leq \|\bar{\mathbf{x}}(t) - \mathbf{v}\|^2 - \frac{2\alpha(t+1)}{n} (F(\bar{\mathbf{x}}(t)) - F(\mathbf{v})) \\ &\quad + \frac{4\alpha(t+1)}{n} \sum_{i=1}^n L_i \|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| + \alpha^2(t+1) \frac{(\mathbf{1}' L_F)^2}{n^2}. \end{aligned}$$

■

With all the pieces in place, we are finally ready to prove Theorem 1. The proof idea is to show that the averages $\bar{\mathbf{x}}(t)$, as defined in Lemma 13, converge to some solution $x^* \in X^*$ and then show that $\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)$ converges to 0 for all i , as $t \rightarrow \infty$. The last step will be accomplished by invoking Lemma 3 on the perturbed push-sum protocol.

Proof of Theorem 1. We begin by observing that the subgradient-push method may be viewed as an instance of the perturbed push-sum protocol. Indeed, let us adopt the notation $\tilde{x}_\ell(t), \tilde{g}_\ell(t)$ from the proof of Lemma 13, and moreover let us define $\tilde{w}_\ell(t), \tilde{z}_\ell(t)$ identically. Then, the definition of subgradient-push implies that for all $\ell = 1, \dots, d$,

$$\begin{aligned} \tilde{w}_\ell(t+1) &= A(t) \tilde{x}_\ell(t), \\ y(t+1) &= A(t) y(t), \\ \tilde{z}_\ell(t+1) &= \frac{\tilde{x}_\ell(t)}{y_\ell(t)}, \\ \tilde{w}_\ell(t+1) &= \tilde{x}_\ell(t+1) - \alpha(t+1) \tilde{g}_\ell(t+1). \end{aligned}$$

Since $\alpha(t) \rightarrow 0$, the assumptions of Lemma 3 are satisfied with $\epsilon(t+1) = \alpha(t+1) \tilde{g}_\ell(t+1)$, from Lemma 3(b) we obtain the conclusion

$$\lim_{t \rightarrow \infty} \left| [\tilde{z}_\ell(t+1)]_i - \frac{\sum_{j=1}^n [\tilde{x}_\ell(t)]_j}{n} \right| = 0 \quad \text{for all } \ell = 1, \dots, d \text{ and all } i = 1, \dots, n,$$

which is equivalent to

$$\lim_{t \rightarrow \infty} \|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| = 0 \quad \text{for all } i = 1, \dots, n. \quad (25)$$

Next, we apply Lemma 3(c). Since the subgradients $\mathbf{g}_i(s)$ are uniformly bounded, and $\{\alpha(t)\}$ is non-increasing and such that $\sum_{t=1}^{\infty} \alpha^2(t) < \infty$, from $\epsilon(t+1) = \alpha(t+1) \tilde{g}_\ell(t+1)$ it follows that for all $i = 1, \dots, n$ and $\ell = 1, \dots, d$,

$$\sum_{t=1}^{\infty} \alpha(t) |\epsilon_i(t+1)| < \sum_{t=1}^{\infty} \alpha(t) \alpha(t+1) \|\tilde{g}_\ell(t+1)\|_\infty \leq \sum_{t=1}^{\infty} \alpha^2(t) \|\mathbf{g}_i(t+1)\|_\infty < \infty.$$

In view of the preceding relation and the assumption that the sequence $\{\alpha(t)\}$ is non-increasing, by applying Lemma 3(b) to each coordinate $\ell = 1, \dots, d$, we obtain

$$\sum_{t=0}^{\infty} \alpha(t+1) \left| [\tilde{z}_\ell(t+1)]_i - \frac{\sum_{j=1}^n [\tilde{x}_\ell(t)]_j}{n} \right| < \infty \quad \text{for all } \ell = 1, \dots, d \text{ and all } i = 1, \dots, n,$$

which implies that

$$\sum_{t=0}^{\infty} \alpha(t+1) \|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| < \infty \quad \text{for all } i = 1, \dots, n. \quad (26)$$

Next we consider Lemma 13 where we use $\mathbf{v} = z^*$ for some solution $z^* \in Z^*$,

$$\begin{aligned} \|\bar{\mathbf{x}}(t+1) - z^*\|^2 &\leq \|\bar{\mathbf{x}}(t) - z^*\|^2 - \frac{2\alpha(t+1)}{n} (F(\bar{\mathbf{x}}(t)) - F^*) \\ &\quad + \frac{4\alpha(t+1)}{n} \sum_{i=1}^n L_i \|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| + \alpha^2(t+1) \frac{(\mathbf{1}' L_F)^2}{n^2}, \end{aligned} \quad (27)$$

where F^* is the optimal value (i.e., $F^* = F(z^*)$ for any $z^* \in Z^*$). In view of Eq. (26), it follows that

$$\sum_{t=1}^{\infty} \frac{4\alpha(t+1)}{n} \sum_{i=1}^n L_i \|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| < \infty.$$

Also, by assumption we have that $\sum_{t=1}^{\infty} \alpha(t) = \infty$ and $\sum_{t=1}^{\infty} \alpha^2(t) < \infty$. Thus, all the conditions of Lemma 12 are satisfied, and by this lemma we conclude that the average sequence $\{\bar{\mathbf{x}}(t)\}$ must converge to some solution $\hat{z} \in Z^*$. By relation (25) it follows that the sequence $\{\mathbf{z}_i(t)\}$, $i = 1, \dots, n$, also converges to the solution \hat{z} . ■

Having proven Theorem 1, we now turn to the proof of the convergence rate results of Theorem 2. The first step will be a slight modification of the result we have just proved - whereas the proof of Theorem 1 shows that $F(\bar{\mathbf{x}}(t)) \rightarrow F(z^*)$, we will now need to argue that we can replace $F(\bar{\mathbf{x}}(t))$ by F evaluated at a running average of the vectors $\bar{\mathbf{x}}(t)$. This is stated precisely in the next lemma.

Lemma 14 *If all the assumptions of Theorem 1 are satisfied and $\alpha(t) = 1/\sqrt{t}$, then for all $t \geq 1$,*

$$\begin{aligned} F\left(\frac{\sum_{k=0}^t \alpha(k+1) \bar{\mathbf{x}}(k)}{\sum_{k=0}^t \alpha(k+1)}\right) - F(z^*) &\leq \frac{n \|\mathbf{x}(0) - \mathbf{z}^*\|_1}{4(\sqrt{t+2}-1)} + \frac{(\sum_{i=1}^n L_i)^2 (1 + \ln t)}{4n(\sqrt{t+2}-1)} \\ &\quad + 8 \frac{(\sum_{i=1}^n L_i) \left(\sum_{j=1}^n \|\mathbf{x}_j(0)\|_1\right) + (\sum_{i=1}^n L_i^2) (1 + \ln t)}{\delta(1-\lambda)(\sqrt{t+2}-1)}. \end{aligned}$$

Proof. From Lemma 13 we have for any \mathbf{v} ,

$$\sum_{k=0}^t \frac{2\alpha(k+1)}{n} (F(\bar{\mathbf{x}}(k)) - F(\mathbf{v})) \leq \|\bar{\mathbf{x}}(0) - \mathbf{v}\|^2 + \sum_{k=0}^t \frac{4\alpha(k+1)}{n} \sum_{i=1}^n L_i \|\mathbf{z}_i(k+1) - \bar{\mathbf{x}}(k)\|$$

$$+ \sum_{k=0}^t \alpha^2(k+1) \frac{(\mathbf{1}' L_F)^2}{n^2}.$$

From the preceding relation, recalling the notation $S(t) = \sum_{k=0}^{t-1} \alpha(k+1)$ and dividing by $(2/n)S(t+1)$, we obtain for any $\mathbf{v} \in \mathbb{R}^d$,

$$\begin{aligned} & \frac{\sum_{k=0}^t \alpha(k+1) F(\bar{\mathbf{x}}(k))}{S(t+1)} - F(\mathbf{v}) \leq \frac{\frac{n}{2} \|\bar{\mathbf{x}}(0) - \mathbf{v}\|^2}{S(t+1)} \\ & + \frac{2}{S(t+1)} \sum_{k=0}^t \alpha(k+1) \sum_{i=1}^n L_i \|\mathbf{z}_i(k+1) - \bar{\mathbf{x}}(k)\| + \frac{1}{S(t+1)} \sum_{k=0}^t \alpha^2(k+1) \frac{(\mathbf{1}' L_F)^2}{2n}. \end{aligned}$$

Now setting $\mathbf{v} = \mathbf{z}^*$ for some $\mathbf{z}^* \in Z^*$ and using the convexity of F , we obtain

$$\begin{aligned} & F\left(\frac{\sum_{k=0}^t \alpha(k+1) \mathbf{x}(k)}{S(t+1)}\right) - F(\mathbf{z}^*) \leq \frac{\sum_{k=0}^t \alpha(k+1) F(\mathbf{x}(k))}{S(t+1)} - F(\mathbf{z}^*) \\ & \leq \frac{\frac{n}{2} \|\bar{\mathbf{x}}(0) - \mathbf{z}^*\|^2}{S(t+1)} + \frac{2}{S(t+1)} \sum_{k=0}^t \alpha(k+1) \sum_{i=1}^n L_i \|\mathbf{z}_i(k+1) - \bar{\mathbf{x}}(k)\| \\ & + \frac{1}{S(t+1)} \sum_{k=0}^t \alpha^2(k+1) \frac{(\mathbf{1}' L_F)^2}{2n}. \end{aligned} \quad (28)$$

We now bound the quantity $\|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\|$ in Eq. (28) by applying Corollary 10 to each component of it; specifically, we will apply Eq. (15) derived in the proof of that corollary. Indeed, observe that all the assumption of that corollary have been assumed to hold; moreover, the constant D in the statement of that corollary equals L_i when the corollary is applied to the ℓ 'th component $\|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\|$. Therefore, adopting our notation of $\tilde{x}_\ell(t)$ and $\tilde{z}_\ell(t)$ as the vectors that stack up the ℓ 'th components of the vectors $\mathbf{x}_i(t), \mathbf{z}_i(t)$, we get that for all $\ell = 1, \dots, d$ and $i = 1, \dots, n$,

$$\sum_{k=0}^t \frac{1}{\sqrt{t+1}} \left| [\tilde{z}_\ell(k+1)]_i - \frac{1}{n} \sum_{j=1}^n [\tilde{x}_\ell(k)]_j \right| \leq \frac{8}{\delta(1-\lambda)} \|\tilde{x}_\ell(0)\|_1 + \frac{8L_i}{\delta} \frac{1 + \ln t}{1-\lambda},$$

which after some elementary algebra implies that

$$\sum_{k=0}^t \alpha(k+1) \sum_{i=1}^n L_i \|\mathbf{z}_i(k+1) - \bar{\mathbf{x}}(k)\|_1 \leq \frac{8 \left((\sum_{i=1}^n L_i) \left(\sum_{j=1}^n \|\mathbf{x}_j(0)\|_1 \right) + (\sum_{i=1}^n L_i^2) (1 + \ln t) \right)}{\delta(1-\lambda)}. \quad (29)$$

Now, using the fact that the Euclidean norm of a vector is not more than its 1-norm, we substitute Eq. (29) into Eq. (28). Then, using the definition of $S(t)$ and Eq. (14) we bound the denominator in Eq. (28) as follows

$$S(t+1) = \sum_{k=0}^t \alpha(k+1) \geq 2 \left(\sqrt{t+2} - 1 \right).$$

The preceding relation and

$$\sum_{k=0}^t \alpha^2(k+1) = \sum_{s=1}^{t+1} \frac{1}{s} \leq 1 + \int_1^t \frac{dx}{x} = 1 + \ln t,$$

yield the stated estimate. ■

We are now finally in position to prove Theorem 2. At this point, the proof is a simple combination of Lemma 14, which tells us that $F\left(\frac{\sum_{k=0}^t \alpha(k+1)\bar{\mathbf{x}}(k)}{\sum_{k=0}^t \alpha(k+1)}\right)$ approaches $F(z^*)$, along with Corollary 10, which tells us that $\frac{\sum_{k=0}^t \alpha(k+1)\bar{\mathbf{x}}(k)}{\sum_{k=0}^t \alpha(k+1)}$ and $\frac{\sum_{k=0}^t \alpha(k+1)\mathbf{z}_i(k+1)}{\sum_{k=0}^t \alpha(k+1)}$ get very close to each other over time for all $i = 1, \dots, n$.

Proof of Theorem 2. It is easy to see by induction that the vectors $\tilde{\mathbf{z}}(t)$ defined in the statement of Theorem 2 are equal to

$$\tilde{\mathbf{z}}(t+1) = \frac{\sum_{k=0}^t \alpha(k+1)\mathbf{z}_i(k+1)}{\sum_{k=0}^t \alpha(k+1)}.$$

By the boundedness of subgradients and Corollary 10, we obtain that

$$F\left(\frac{\sum_{k=0}^t \alpha(k+1)\bar{\mathbf{x}}(k)}{\sum_{k=0}^t \alpha(k+1)}\right) - F(\tilde{\mathbf{z}}(t+1)) \leq 4 \frac{(\sum_{i=1}^n L_i) \left(\sum_{j=1}^n \|\mathbf{x}_j(0)\|_1\right) + (\sum_{i=1}^n L_i^2) (1 + \ln t)}{\delta(1-\lambda)(\sqrt{t+2}-1)}$$

Finally, Lemma 14 along with the inequality $2(\sqrt{t+2}-1) \geq \sqrt{t}$ now implies Theorem 2. ■

5 Conclusions

We have introduced the subgradient-push, a broadcast-based distributed protocol for distributed optimization of a sum of convex functions over directed graphs. We have shown that, as long as the communication graph sequence $\{G(t)\}$ is uniformly strongly connected, the subgradient-push succeeds in driving all the nodes to the same point in the set of optimal solutions. Moreover, the objective function converges at a rate of $O(\ln t/\sqrt{t})$, where the constant depends on the initial vectors, bounds on subgradient norms, consensus speed λ of the graph sequence $\{G(t)\}$, as well as a measure of the imbalance of influence δ among the nodes.

Our results motivate the open problems associated with understanding how the consensus speed λ depends on properties of the sequence $\{G(t)\}$. Similarly, it is also natural to ask how the measure of imbalance of influence δ depends on the combinatorial properties of the graphs $G(t)$, namely how it depends on the diameters, the size of the smallest cuts, and other pertinent features of the graphs in the sequence $\{G(t)\}$.

References

- [1] F. Benezit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli. Weighted gossip: distributed averaging using non-doubly stochastic matrices. In *Proceedings of the 2010 IEEE International Symposium on Information Theory*, Jun. 2010.

- [2] V.D. Blondel, J.M. Hendrickx, A. Olshevsky, and J.N. Tsitsiklis. Convergence in multi-agent coordination, consensus, and flocking. In *Proceedings of the IEEE Conference on Decision and Control*, Dec 2005.
- [3] J. Chen and A. H. Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, 2012.
- [4] A.D. Dominguez-Garcia and C. Hadjicostis. Distributed strategies for average consensus in directed graphs. In *Proceedings of the IEEE Conference on Decision and Control*, Dec 2011.
- [5] J.C. Duchi, A. Agarwal, and M.J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, March 2012.
- [6] B. Gharesifard and J. Cortes. Distributed continuous-time convex optimization on weight-balanced digraphs. <http://arxiv.org/pdf/1204.0304.pdf>, 2012.
- [7] A. Jadbabaie, J. Lin, and A.S. Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48:988–1001, 2003.
- [8] B. Johansson. *On distributed optimization in networked systems*. PhD thesis, Royal Institute of Technology (KTH), tRITA-EE 2008:065, 2008.
- [9] D. Kempe, A. Dobra, and J. Gehrke. Gossip-based computation of aggregate information. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 482 – 491, Oct. 2003.
- [10] H. Li and Z. Han. Competitive spectrum access in cognitive radio networks: graphical game and learning. In *IEEE Wireless Communications and Networking Conference*, pages 1–6, 2010.
- [11] I. Lobel and A. Ozdaglar. Distributed subgradient methods for convex optimization over random networks. *IEEE Transactions on Automatic Control*, 56(6):1291–1306, June 2011.
- [12] I. Lobel, A. Ozdaglar, and D. Feijer. Distributed multi-agent optimization with state-dependent communication. *Mathematical Programming*, 129(2):255–284, 2011.
- [13] C. Lopes and A. H. Sayed. Incremental adaptive strategies over distributed networks. *IEEE Transactions on Signal Processing*, 55(8):4046–4077, 2007.
- [14] L. Moreau. Stability of multiagent systems with time-dependent communication links. *IEEE Transactions on Automatic Control*, 50:169–182, 2005.
- [15] A. Nedić. Asynchronous broadcast-based convex optimization over a network. *IEEE Transactions on Automatic Control*, 56(6):1337–1351, 2011.

- [16] A. Nedić, A. Olshevsky, A. Ozdaglar, and J.N. Tsitsiklis. On distributed averaging algorithms and quantization effects. *IEEE Transactions on Automatic Control*, 54(11):2506–2517, 2009.
- [17] A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, Jan. 2009.
- [18] A. Nedić and A. Ozdaglar. Convergence Rate for Consensus with Delays. *Journal of Global Optimization*, 47:437456, 2010.
- [19] G. Neglia, G. Reina, and S. Alouf. Distributed gradient optimization for epidemic routing: A preliminary evaluation. In *IEEE Wireless Days, 2nd IFIP*, pages 1–6, 2009.
- [20] A. Olshevsky. *Efficient information aggregation for distributed control and signal processing*. PhD thesis, MIT, 2010.
- [21] B. Polyak. *Introduction to optimization*. Optimization software, Inc., Publications division, New York, 1987.
- [22] M. Rabbat and R. D. Nowak. Distributed optimization in sensor networks. In *IPSN*, pages 20–27, 2004.
- [23] S. S. Ram, A. Nedić, and V. V. Veeravalli. Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization. *Journal of Optimization Theory and Applications*, 147:516–545, 2010.
- [24] S. S. Ram, A. Nedić, and V. V. Veeravalli. A new class of distributed optimization algorithms: application to regression of distributed data. *Optimization Methods and Software*, 27(1):71–88, 2012.
- [25] S. S. Ram, V. V. Veeravalli, and A. Nedić. Distributed non-autonomous power control through distributed convex optimization. In *IEEE INFOCOM*, pages 3001–3005, 2009.
- [26] H. Robbins and D. Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. *Optimizing Methods in Statistics, Academic Press, New York*, pages 233 – 257, 1971.
- [27] K. Srivastava and A. Nedić. Distributed asynchronous constrained stochastic optimization. *IEEE J. Sel. Topics. Signal Process.*, 5(4):772–790, 2011.
- [28] K.I. Tsianos, S. Lawlor, and M.G. Rabbat. Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning. In *Proceedings of the 50th Allerton Conference on Communication, Control, and Computing*, 2012.
- [29] K.I. Tsianos, S. Lawlor, and M.G. Rabbat. Push-sum distributed dual averaging for convex optimization. In *Proceedings of the IEEE Conference on Decision and Control*, 2012.

- [30] J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31:803–812, 1986.